Estimating biosignals using the human voice

Eduardo Coutinho and Björn Schuller*

Department of Computing, Imperial College London, United Kingdom.

*Corresponding Author: bjoern.schuller@imperial.ac.uk

Computational paralinguistics (CP) is a relatively new area of research that provides new methods, tools, and techniques to automatically recognize the states, traits, and qualities embedded in the nonsemantic aspects of human speech (*1*). In recent years, CP has reached a level of maturity that has permitted the development of a myriad of applications in everyday life, such as the automatic estimation of a speaker's age, gender, height, emotional state, cognitive load, personality traits, likability, intelligibility, and medical condition (*2*). Here, we provide an overview of one particular application of CP that offers new solutions for health care—the recognition of physiological parameters (biosignals) from the voice alone.

Unintrusive and pervasive monitoring

Currently, there are a variety of portable medical devices enabling patients to actively monitor the relevant factors contributing to their diagnosis and treatments. These devices are particularly important when frequent monitoring (daily or several times a day) is required for the adequate treatment and detection of symptoms, especially for patients with limited mobility and difficulties accessing medical facilities. Further, these technologies help address the shortage of qualified medical staff needed to adequately monitor patients, which can lead to delays in obtaining appropriate feedback and treatment.

The technologies currently available include those that measure heart rate, blood volume pressure, body temperature, respiration rate, and other physiological parameters. Such devices can be quite expensive and complicated for older patients and those with limited mobility, and often inconvenient for everyday use. Ideally, monitoring biosignals should be unobtrusive, not require additional electronic devices, and require minimal effort from the patient. Most importantly, monitoring should be easy to perform in emergency situations.

Computers or mobile phones are thus an obvious choice due to their abundance and their computational power, which is sufficient to acquire and analyze biosignals (*3–5*). If such devices are to be used, the signal being measured must be one that can be recorded without the need for additional equipment. Audio and video signals fit these criteria, as both have been previously used to estimate a variety of biosignals. For instance, video analysis of the skin can detect subtle color shifts triggered by physiological changes (such as cardiac rhythm or blood flow) (*6–8*). In the case of the human voice, physiological changes are detectable through vocalizations because both the larynx (where the vocal cords are located) and the pharynx (above the larynx) are controlled by the autonomic nervous system, which regulates blood pressure, heart rate, and perspiration (*9–12*).

Voice-based biosignal estimation presents a major advantage over video-based sensing, because audio acquisition is less limiting than video in that it does not need to be directed toward

or be in contact with a patient's skin, and it can be used in a wider range of conditions (for instance, in the dark when video cannot be captured). This is of particular relevance in crisis situations, when additional sensors or the ideal conditions for adequate video analysis are not available. In such cases, by simply asking for medical assistance, vital information about the patient could be automatically collected and used to inform diagnosis and treatment.

Voice-based physiological monitoring

In a recent and comprehensive attempt to estimate biosignals from the voice alone (*13*), we evaluated the estimation of two biosignals—heart rate (HR) and skin conductance (SC)—and the classification of pulse level (high pulse/low pulse; HP/LP) using acoustic features extracted from audio recordings. We designed an empirical study to collect subjects' HR and SC from 19 subjects (4 female; 15 male). In addition, we obtained audio recordings of breathing sounds and from the repeated pronunciation of the sustained vowel "a." The recordings were collected in two pulse-level states: a "neutral" state (characterized by a low pulse), and a high-pulse state, which was induced by asking subjects to run up and down six flights of stairs (three stories) and down a hallway immediately prior to the recording. In order to evaluate the influence of the sound recording conditions, audio recordings were obtained with two different devices: a high-quality sound recorder ("ambient") and a common, commercially available headset ("headset"). The full database consists of 1,420 audio recordings (and concomitant HR and SC recordings).

The audio recordings were analyzed using the openSMILE (Speech and Music Interpretation by Large Space Extraction) software toolkit (*14*), which was used to extract a large set of acoustic descriptors. These descriptors included 4,368 acoustic features comprising a variety of energy-, spectral-, and voice-related information, which was used to develop computational models that predict SC, HR, and pulse level using state-of-the-art machine learning regression (to determine the exact value of SC and HR) and classification techniques (to determine pulse level, either high or low). These computational models were created for individual speakers (IS) and multiple speakers (MS), i.e., using the recordings from all speakers to generate a model that can be applicable to any speaker rather than a specific speaker. The models' performances in the regression tasks were estimated using Pearson's linear correlation coefficient (CC) and the mean absolute error (MAE). For the classification of pulse levels, the performance was estimated using the unweighted accuracy [UA, i.e., the unweighted arithmetic mean of the number of correctly identified pulse levels in each condition (HP or LP)]. A summary of the results is shown in Table 1.

Next, we evaluated whether voice recordings could be used to identify pulse level and estimate SC and HR values. The results demonstrated that one's pulse level could be correctly identified with an accuracy as high as 84.1 percent or the IS model (ambient microphone audio recordings of breathing). Furthermore, HR and SC regression analysis showed that the best linear correlation coefficients were 0.861 [MAE of 8.1 beats per minute (BPM); IS model using the sustained vowels audio recordings obtained with an ambient microphone] and 0.978 [MAE of 84.4 micromhos (μMhO); IS model using the sustained vowels audio recordings obtained with a headset microphone)], respectively. We drew three main conclusions from the results. First, common microphones are a viable option for estimating biosignals from the voice, as the performance was comparable for both microphone types. Secondly, both types of recording conditions—sustained vowels and breathing sounds—led to similar classifications of the subjects' pulse level, although the use of sustained vowels was slightly better than breathing sounds for the regression experiments (*13*). In another study, we evaluated which acoustic

features would be best suited for estimating biosignals. The results showed that with an optimized set of 150 acoustic features, a subject's pulse level could be accurately determined, with a UA of 91.4 percent and correlation coefficients of 0.876 for SC and 0.838 for HR (but only when using 100 acoustic features for the analysis, not 150) (*15*).

The dataset used in our work—the Munich BioVoice corpus (MBC)—has been made publicly available (*15*) and was used in the Interspeech 2014 Computational Paralinguistics Challenge (*2*). Competing teams were asked to classify HP/LP based on freely chosen features extracted from voice recordings of text that was read after exercise or under resting conditions. The winning team achieved an accuracy of 75.3 percent (*16*).

Conclusions and perspectives

Taken together, our studies have shown that audio recordings of a person's breathing, pronunciation of sustained vowels, and reading of text can be used to predict biosignals. Gathering such information from voice recordings has a lot of potential use for technologies that require noninvasive, passive collection methods. For instance, a mobile phone could be used to continuously or periodically record a subject's voice (with or without speech) without the need for user intervention. This would require little or no effort from the user and be well suited to patients with limited mobility or in emergency situations when user intervention is not possible.

However, the use of audio recordings to estimate biosignals is still in its infancy and has a lot of room for improvement. For example, the data from this technology is still less accurate than what can be obtained by using dedicated medical equipment, and more research is needed to improve its quality. Furthermore, the technology would gain from research on which acoustic and vocal features are optimal to use, from exploration of more powerful modeling paradigms, from the calibration of models based on individual differences in physiological activity, and from the acquisition of larger data sets for refining speaker-independent models. Finally, this type of research would benefit from more attention from the speech community and from the application of state-of-the-art machine learning techniques.

In summary, our studies have found that audio-based recognition has the potential to be used as a software application on mobile phones and computers for remote monitoring of HR and SC. One advantage of using such audio recordings is that analyses could be performed in an atemporal fashion, e.g., using past recordings to investigate a patient's history and their condition's evolution over the period that preceded diagnosis and treatment. If the technology is further improved, it could be used for passive, noncontact monitoring of patients, which would require minimum attendance by its user and improve the quality of life for many people.

References

1. B. Schuller, A. Batliner, *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing* (Wiley, New York, 2013).

2. B. Schuller *et al.*, in *Proceedings of the 15th Annual Conference of the International Speech Communication Association* (ISCA, Dresden, 2014), pp. 148–152.

3. M. N. Boulos *et al*., *Biomed. Eng. Online* **10** (2011).

4. E. Kyriacou, C. Pattichis, M. Pattichis, in *31st Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (IEEE, Minneapolis, MN, 2009), pp. 1246–1249.

5. C. G. Scully *et al*., *IEEE Trans. Biomed. Eng.* **59**, 303 (2012).

6. M.-Z. Poh, D. J. McDuff, R. W. Picard, *IEEE Trans. Biomed. Eng.* **58**, 7 (2011).

7. E. Jonathan, M. J. Leahy, *Physiol. Measurement* **31**, 79 (2010).

8. E. Jonathan, M. J. Leahy, *J. Biophotonics* **4**, 293 (2011).

9. R. F. Orlikoff, R. Baken, *J. Speech, Hear. Res.* **32**, 576 (1989).

10. Saloni, R. K. Sharma, A. K. Gupta, *Int. J. Image, Graphics and Signal Process.* **6**, 47 (2014).

11. A. Mesleh, D. Skopin, S. Baglikov, A. Quteishat, *J. Comput. Sci. Tech.* **27**, 1243 (2012).

12. D. Skopin, S. Baglikov, in *Proceedings of the 4th International Conference on Information Technology* (2009).

13. B. Schuller, F. Friedmann, F. Eyben, in *Proceedings of the 38th IEEE International Conference on Acoustic, Speech, and Signal Processing* (IEEE, Vancouver, 2013), pp. 7219–7223.

14. F. Eyben, F. Weninger, F. Groß, B. Schuller, in *Proceedings of the ACM International Conference on Multimedia* (ACM, Barcelona, Spain, 2013), pp. 835– 838.

15. B. Schuller, F. Friedmann, F. Eyben, in *Proceedings of the Language Resources and Evaluation Conference* (ELRA, Reykjavik, 2014), pp. 1506–1510.

16. H. Kaya, T. Özkaptan, A. A. Salah, S. F. Gürgen, in *Proceedings of the 15th Annual Conference of the International Speech Communication Association,* pp. 442–446 (ISCA, Dresden, 2014).

Acknowledgments This work was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement 645378 [Artificial Retrieval of Information Assistants-Virtual Agents with Linguistic Understanding, Social Skills, and Personalised Aspects (ARIA-VALUSPA)].

Table 1. Results for the automatic regression of heart rate (HR), skin conductance (SC), and classification of pulse level (HP: high pulse; LP: low pulse). IS: individual speakers; MS: multiple speakers; UA: unweighted accuracy; CC: Pearson's linear correlation coefficient; MAE: mean absolute error. Table adapted from (*13*).

Recording condition	**Recording device**	**Model type**	**Pulse level (HP/LP)**	**Heart Rate (HR)**		**Skin Conductance (SC)**	
			UA (%)	CC	MAE (BPM)	CC	MAE (µMhO)
Sustained vowels	Headset	IS	83.1	0.809	8.4	0.978	84.4
		MS	79.6	0.770	10.6	0.891	265.3
	Ambient	IS	82.7	0.861	8.1	0.960	88.2
		MS	76.0	0.574	11.7	0.633	311.2
Breathing periods	Headset	IS	84.1	0.722	10.7	0.908	153.7
		MS	78.6	0.629	13.1	0.632	469.7
	Ambient	IS	81.9	0.718	10.6	0.905	165.3
		MS	72.9	0.521	14.8	0.483	570.8